Project 1
-
SNP Population Analysis of Chromosome 3

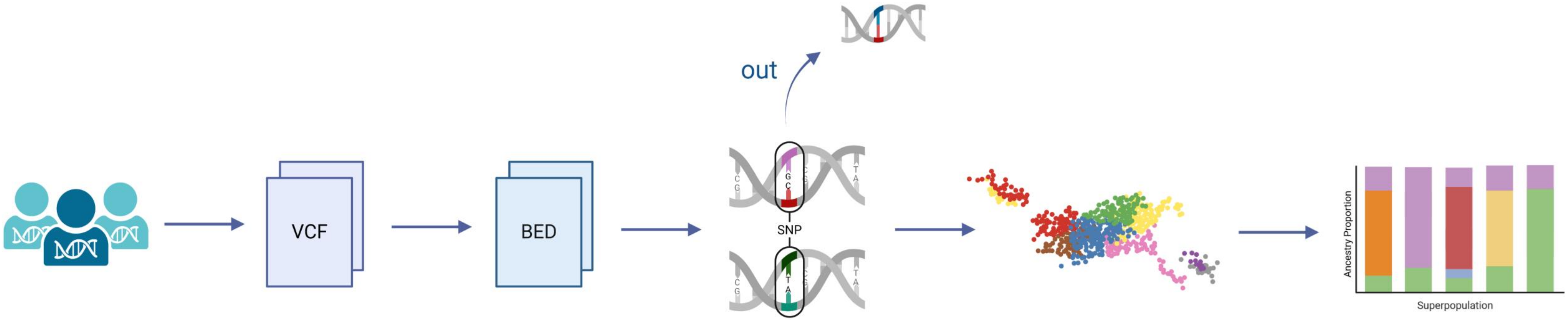
Sara Burckhardt
Majorie Bärtschi
BIO392

Goal of the Project

To analyze genetic population structure and ancestral components of individuals using genome-wide SNP data from chromosome 3.

What are the ancestry proportions for individuals in a superpopulation? Do individuals from one superpopulation nicely cluster together?

Methods - Workflow



Created with bioRender

Data from 1000 Genome Project

File Conversion with PLINK

Filtering and Pruning

PCA

Admixture Analysis

Methods and Tools

What is PLINK?

- PLINK is a tool that is widely used in population genetics for file conversion, quality control, filtering and analysis of large-scale genotype data.

Why Pruning and filtering?

- We filter and prune to remove low-quality SNPs, correlated SNPs or SNPs which are not so rare, meaning they have a Minor Allele Frequency (MAF) of more than **5%.**

What is needed for PCA?

- The files from the Pruning were used in PLINK to run the PCA which creates two files (.eigenvec and .eigenval) containing the PC scores and Eigenvalues (amount of variance explained).

What does Admixture Analysis tell us?

- Admixture analysis looks at how much of an individual’s genome is derived from different ancestral populations. With that we gain information about genetic ancestry.

Why is Linkage Disequilibrium Pruning important?

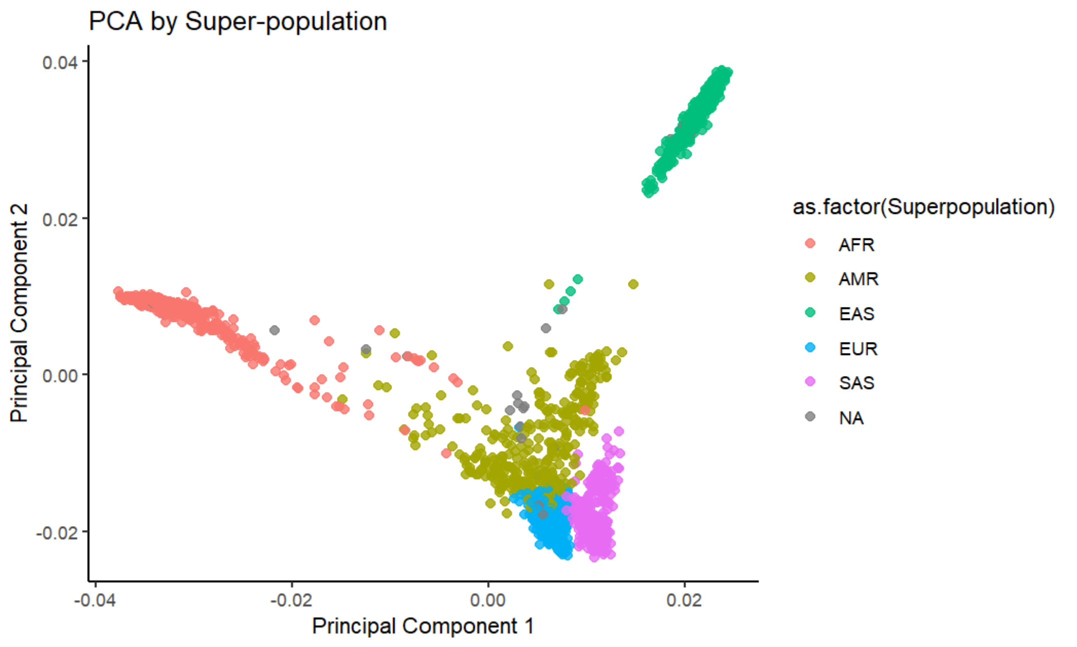
Tools like Admixture or PCA assume that SNPs are unlinked. Therefore, we need Linkage Disequilibrium Pruning to ensure we remove correlated SNPs. How do we filter out SNPs that are correlated?

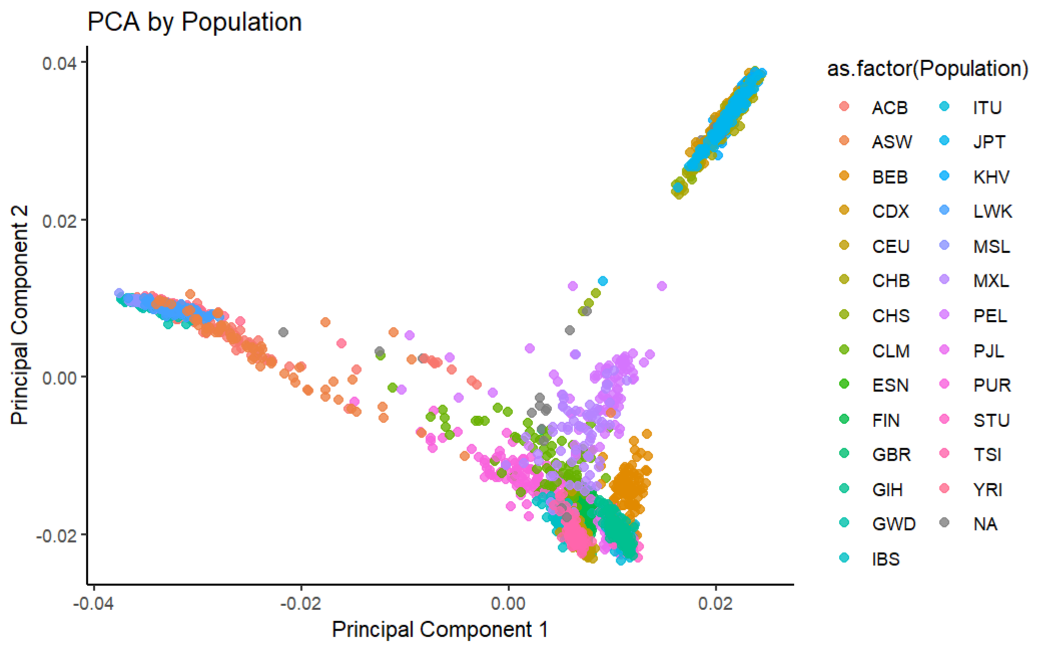
r^2 = 1 → two SNPs are perfectly correlated (always inherited together)

r^2 = 0 → two SNPs are completely independent (randomly inherited)

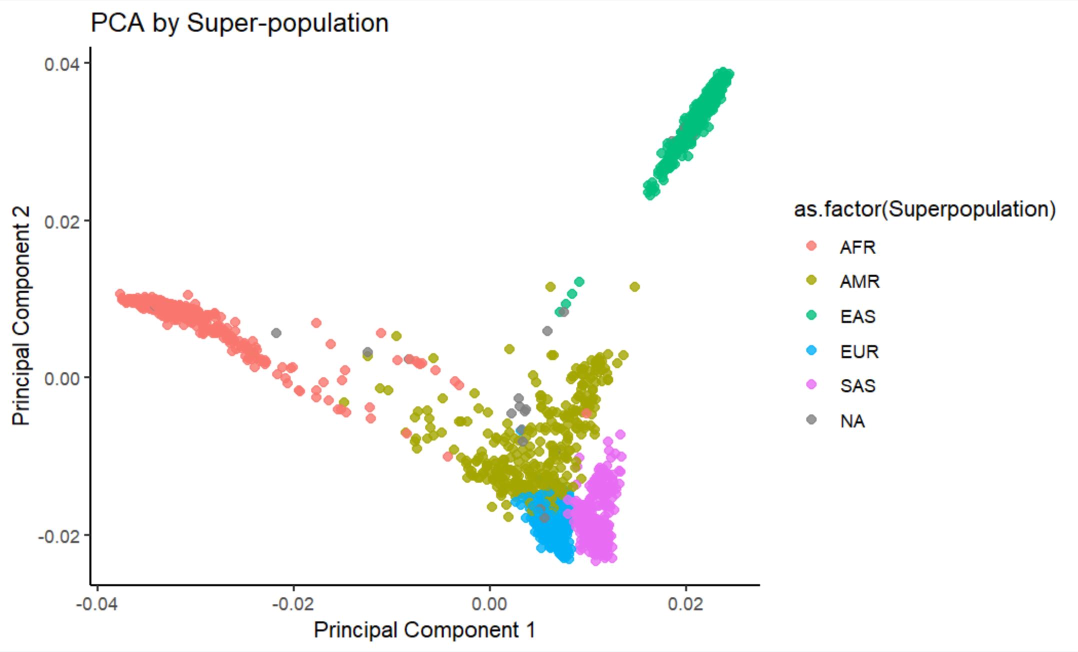
Our **threshold was r^2 = 0.2**, so all pairs of SNPs that are above that value are filtered out.

Results - PCA

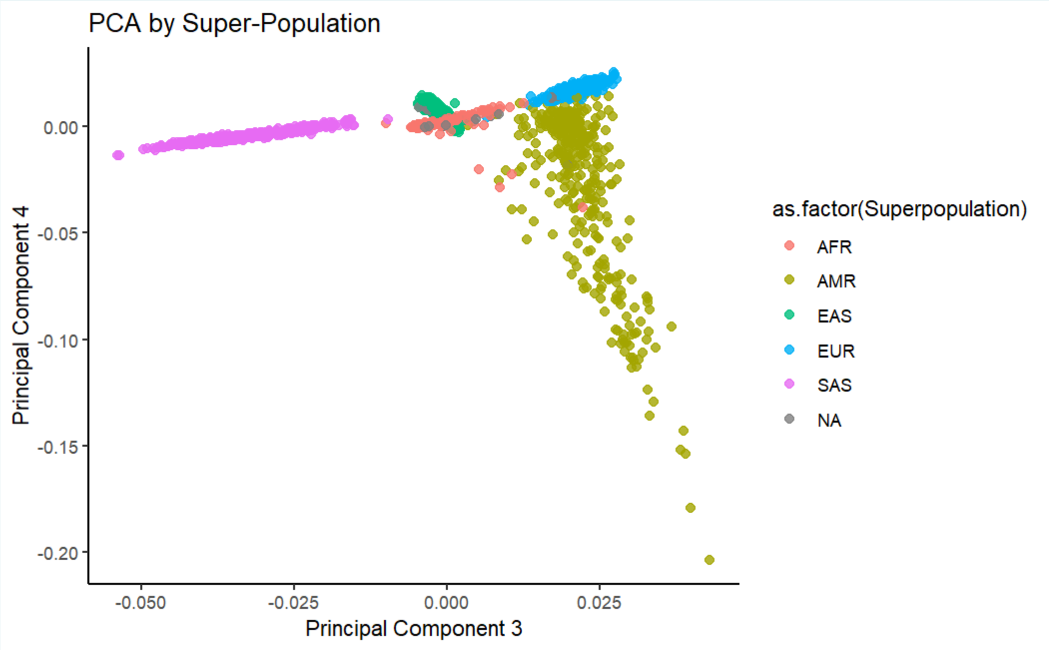

PCA by Super-population


PCA by Population

Results – PCA

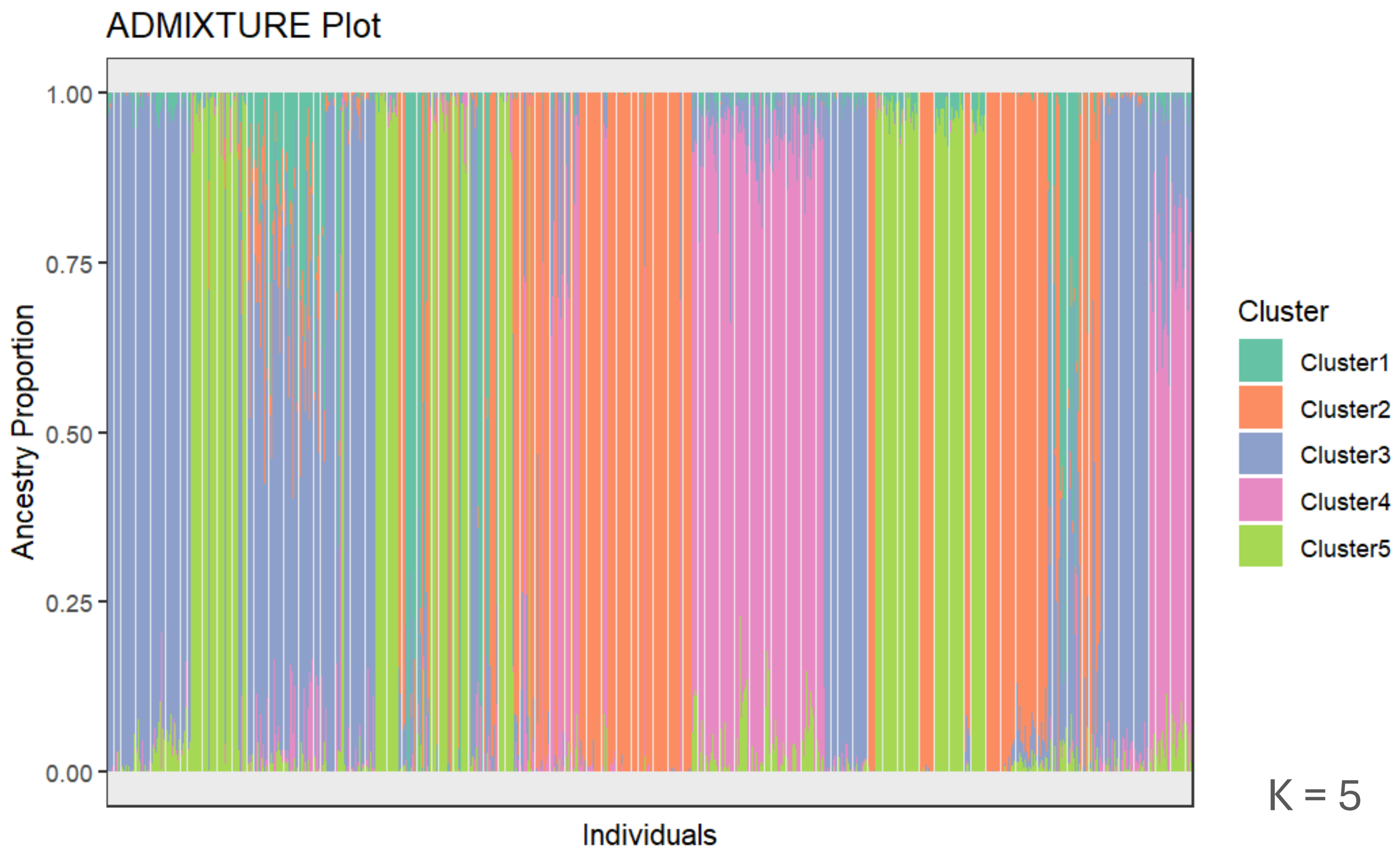


PCA by Super-population



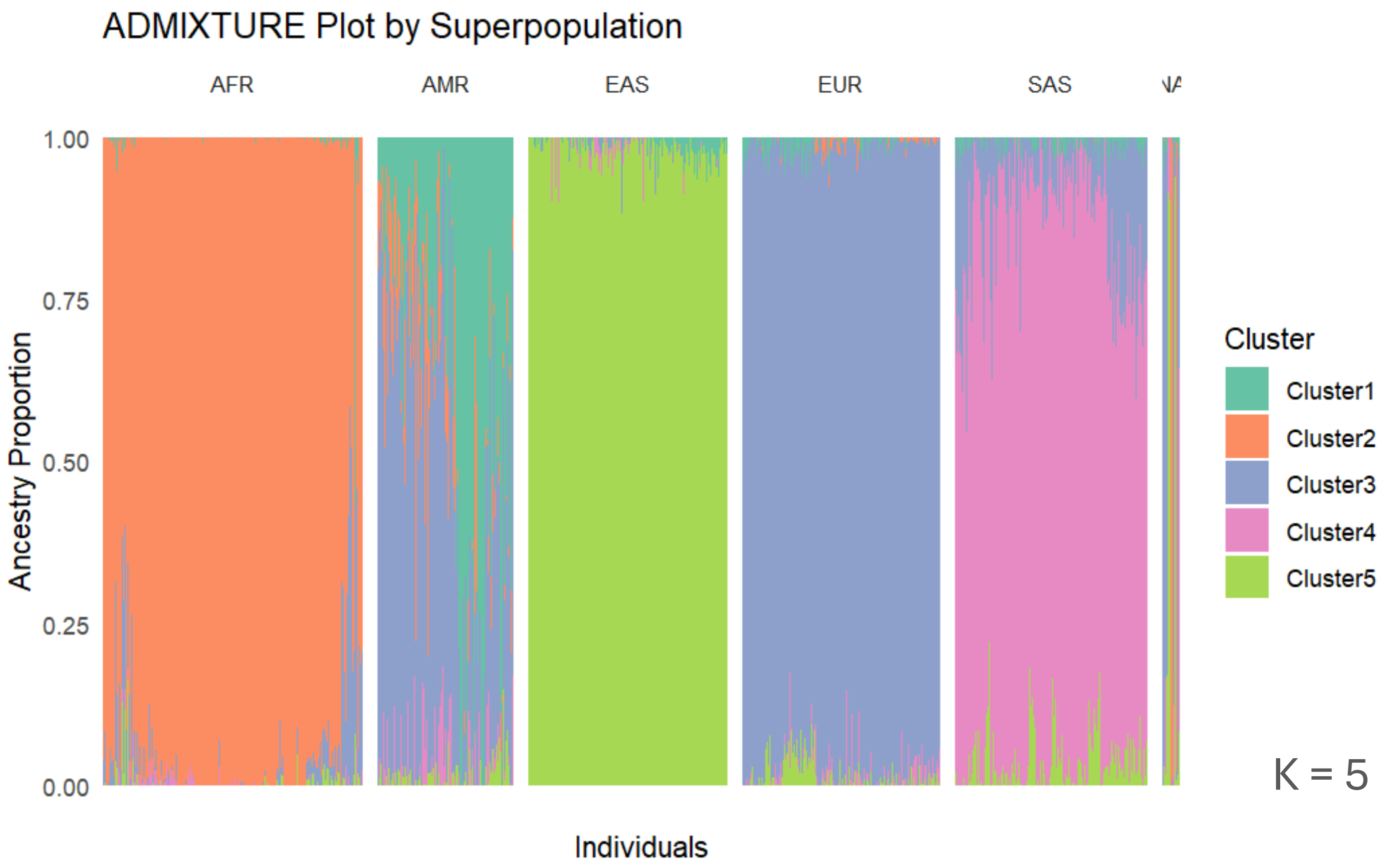
PCA by Super-Population

Results – Admixture Analysis



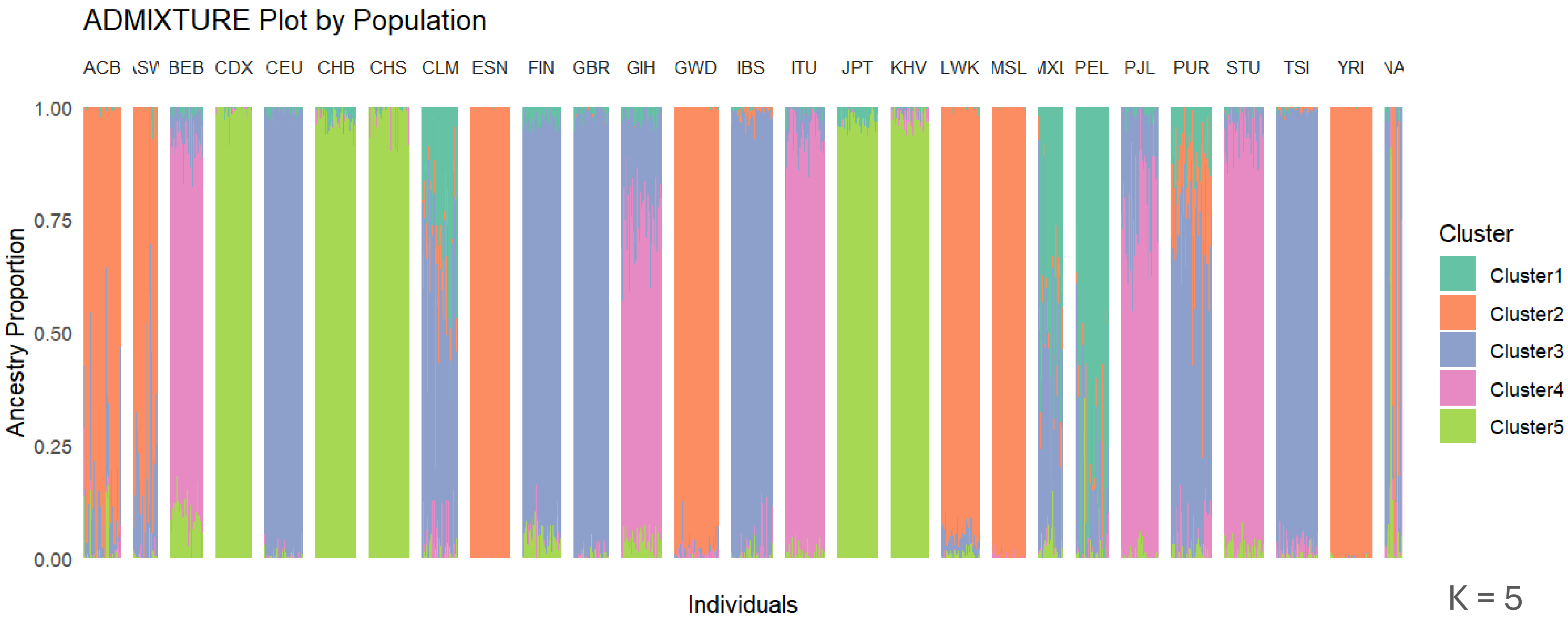
K = 5

Results – Admixture Analysis



K = 5

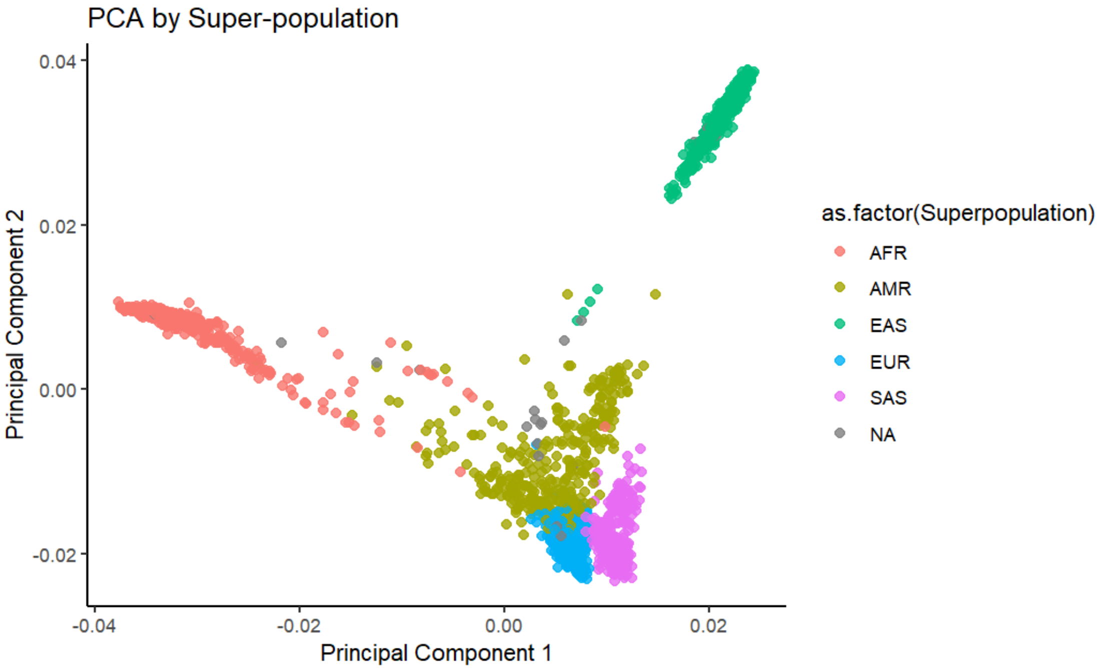
Results – Admixture Analysis

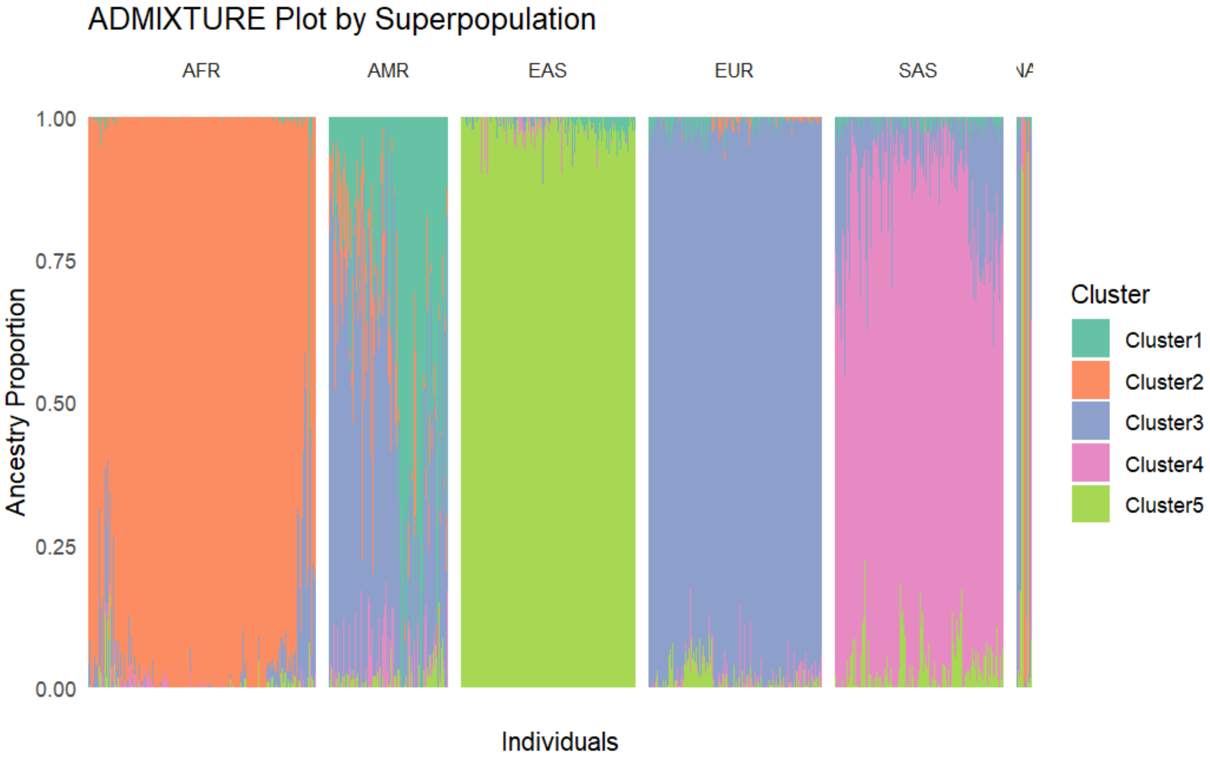


K = 5

Results - Comparison

Do PCA and Admixture show similar results?





Summary / Conclusion

What we found: We see clear population structure, meaning individuals from one superpopulation are predominantly assigned to one cluster.

Next steps / Further Analysis: Admixture analysis with different K-sizes, to find the K that best explains genetic variation that is observed.

Thank you!

-

Questions?

References

Datasets

- https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20181203_biallelic_SNV/
- https://drive.google.com/drive/folders/1fEy09eRa0Cs4O_paZvyO5rAwnfvdt7M-?usp=sharing

Reading

- https://link.springer.com/protocol/10.1007/978-1-0716-0199-0_4
- https://connor-french.github.io/intro-pop-structure-r/

- Tools and Programs:
- https://www.cog-genomics.org/plink/
- https://dalexander.github.io/admixture/download.html